DOCUMENT RESUME

ED 403 301 TM 026 025

AUTHOR Thompson, Bruce
TITLE The Treatment of Score Reliability and Validity in the New ANSI-Approved Program Evaluation Standards.
PUB DATE Aug 96
NOTE 20p.; Paper presented at the Annual Meeting of the American Psychological Association (104th, Toronto, Canada, August 1996).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Program Evaluation; Psychometrics; *Reliability; *Scores; *Standards; Test Interpretation; *Validity
IDENTIFIERS American Educational Research Association; *American National Standards Institute; American Psychological Association; National Council on Measurement in Education

ABSTRACT

The program evaluation standards approved by the American National Standards Institute (ANSI) in 1994 that deal with reliability and validity accurately represent contemporary views of the psychometric community with regard to reliability and validity. As such, these standards move the field forward. The ANSI standards recognize that reliability is a characteristic of the scores or of the data in hand, rather than being a characteristic of the test. It is also essential to recognize, as the ANSI standards do, that it is the inferences that are made from scores, rather than the test, that are valid. It is to be hoped that the next edition of the standards developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education will provide approximately the same treatment of reliability and validity. An appendix gives a suggested addition to the 1996 standards. (Contains 31 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

apajoint.wp1 7/10/96

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
BRUCE THOMPSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 403 301

The Treatment of Score Reliability and Validity in the New ANSI-Approved Program Evaluation Standards

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada, August, 1996.

TM026025

BEST COPY AVAILABLE

ABSTRACT

The 1994 ANSI-approved program evaluation standards dealing with reliability and validity accurately represent contemporary views of reliability and validity within the psychometric community. As such, these standards (Joint Committee, 1994) move the field forward. One can only hope that the next in a series of editions of the AERA/APA/NCME standards will provide approximately the same treatment of reliability and validity.

Program evaluation has had important influences on educational and other interventions during the last several decades (Thompson, 1994b). The new standards for program evaluation, developed by the Joint Committee on Standards for Educational Evaluation (1994), were recently published. These standards mark yet another milestone in the use of program evaluation in service of improving programs.

The Joint Committee itself has had important impacts, reflected in its history.

- The Joint Committee first met in 1975. The Joint Committee consists of representatives of various organizations and at-large members. In 1981 the Joint Committee was incorporated as a non-profit organization. Currently, the Joint Committee consists of representatives of 15 organizations (e.g., AFT, NEA, AERA, NCME, APA, NAESP, NASP, ASCD, CCSSO, NSBA).
- In 1981 the Joint Committee's *standards for program evaluation* were published.
- In 1988 the Joint Committee's *standards for personnel evaluation* were published (Joint Committee, 1988).
- In 1989 the procedures of the Joint Committee's were certified by the American National Standards Institute (ANSI). Standards approved by ANSI become <u>the</u> approved American national standards.
- In 1994 the Joint Committee's revised *standards for program evaluation* were published and became the first ANSI-approved standards focusing on professional principles and conduct.

Sanders (1994) elaborated the process of developing professional standards that are ANSI-approved.

The purpose of the present paper is to review the treatment, within the 1994 ANSI-certified program evaluation standards, of issues involving score reliability and score validity. The analysis includes consideration of recent trends in thinking about both score reliability and score validity, and some comparisons with the treatment of these issues in the AERA/APA/NCME (1985) test standards.

Score Reliability

Tests are NOT Reliable

Too few researchers act on a conscious recognition that *reliability is a characteristic of scores or the data in hand*. Test booklets are not impregnated with reliability during the printing process. The same WISC-R that yields reliable scores for some adults on a given occasion of measurement will not necessarily do so when the same test is administered to first-graders.

Many researchers recognize these dynamics on some level, but paradigm influences constrain some researchers from actively integrating this presumption into their actual analytic practice. The pernicious practice of saying, "the test is reliable", creates a language that predisposes researchers against acting on a conscious realization that tests themselves are not reliable, and acting accordingly (Thompson, 1994a).

As Rowley (1976, p. 53, emphasis added) argued, "It needs to be established that an instrument itself is neither reliable nor

unreliable.... A single instrument can produce *scores* which are reliable, and other *scores* which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) argued that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the *scores* on a test for a *particular* group of examinees."

In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

And Eason (1991, p. 84, emphasis added) argued that:

> Though some practitioners of the classical measurement paradigm [incorrectly] speak of reliability as a characteristic of tests, in fact reliability is a characteristic of *data*, albeit data generated on a given measure administered with a given protocol to given subjects on given occasions.

The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of the measurement protocol. Reliability is driven by variance--typically, greater score variance leads to greater score

reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of subjects, will yield scores with differing reliability. As Dawis (1987, p. 486) observed, "...Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

Our shorthand ways of speaking (e.g., language saying "the test is reliable") can itself cause confusion and lead to bad practice. As Pedhazur and Schmelkin (1991, p. 82, emphasis in original) observed, "Statements about *the* reliability of a measure are... inappropriate and potentially misleading." These telegraphic ways of speaking are not inherently problematic, but they often later become so when we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our jargon is sometimes telegraphic and is <u>not</u> literally true. As noted elsewhere:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (Thompson, 1992b, p. 436)

One sloppy practice is not calculating, reporting, and interpreting the reliability of one's own scores for one's own

data. As Pedhazur and Schmelkin (1991, p. 86, emphasis in original) argued:

> Researchers who bother at all to report reliability estimates for the instruments they use (many do not) frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. Such information may be useful for comparative purposes, but it is imperative to recognize that the *relevant reliability estimate is the one obtained for the sample used in the [present] study under consideration*.

Why Score Reliability is So Important

In one book exploring the intimate linkages between measurement error variance and our attributions about the origins of variance in our substantive basic or applied research, Pedhazur and Schmelkin (1991) noted,

> Measurement error is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures... (pp. 2-3)

Therefore, it should not be surprising that studies of research reports in journals indicate insufficient attention is paid to the impacts of measurement integrity on the integrity of substantive research conclusions. For example, with respect to the American Educational Research Journal, Willson (1980) reported that:

> Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research.... That reliability... is unreported in almost half the published research is... inexcusable at this late date.... (pp. 8-9)

A more recent "perusal of contemporary psychology journals demonstrates that quantitative reports of scale reliability and validity estimates are often missing or incomplete" (Meier & Davis, 1990, p. 113); and that "the majority [95%, 85% and 60%] of the scales described in the [three Journal of Counseling Psychology] JCP volumes [1967, 1977 and 1987] were not accompanied by reports of psychometric properties" (p. 115).

This state of affairs is surprising, given two related trends within the literature. First, since the influential articles by Cohen (1968) and Knapp (1978) appeared, more researchers have recognized that all parametric statistical analyses are correlational (Thompson, 1991a), and that substantive variance-accounted-for effect sizes expressed as r^2 analogs can be

interpreted in all studies. Second, the importance of interpreting effect sizes as against statistical significance tests has been increasingly recognized (e.g., Thompson, 1993, 1996), as reflected, for example, in a recent procession of articles within the American Psychologist (cf. Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989).

Nevertheless, too few researchers act on the premise that score reliability establishes a ceiling for substantive effect sizes. These impacts can be readily illustrated in a concrete example using the bivariate correlation as an heuristic.

It has been recognized in textbooks dating back to the 1950s, and in more recent books as well (e.g., Pedhazur & Schmelkin, 1991, p. 114), that a correlation coefficient "corrected" for attenuation due to measurement error ($\hat{r}_{XY}$) can be estimated as:

$$\hat{r}_{XY} = r_{XY} / [(r_{XX} * r_{YY})^{.5}],$$

where r_{XY} is the calculated bivariate relationship between scores on variables $\underline{X}$ and $\underline{Y}$, and r_{XX} and r_{YY} are respectively the reliability coefficients for scores on $\underline{X}$ and $\underline{Y}$. This algorithm can be re-expressed in the more familiar metric of common variance, as is often done in popular variance-accounted-for effect size statistics (e.g., $\underline{r}^2$, R^2, eta^2, $omega^2$):

$$\hat{r}_{XY}^2 = r_{XY}^2 / (r_{XX} * r_{YY})$$

Through algebraic manipulation, the detectable effect size, given knowledge of "true" relationship, $\hat{r}_{XY}^2$, and the reliabilities of the two sets of scores, is:

$$r_{XY}^2 = \hat{r}_{XY}^2 * (r_{XX} * r_{YY})$$

Even if the "true" relationship between perfectly reliable measures of X and Y was perfect, i.e., $\hat{r}_{XY}^2 = 1.0$, the detectable effect in any study can *never exceed the product of the reliability coefficients for the two sets of scores*:

$$r_{XY}^2 = 1 * (r_{XX} * r_{YY})$$

For example, even when $\hat{r}_{XY}^2 = 1.0$, if both sets of scores have reliability coefficients of .7, the detectable effect cannot exceed .49. Clearly, measurement error prospectively impacts the effect size that we can obtain in a planned study and also should be retrospectively considered when interpreting calculated effects once the study has been done.

The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. We may conduct studies that could not possibly yield noteworthy effect sizes. Or we may not accurately interpret our results if we do not consider the reliability of the scores we are actually analyzing.

These practices may be caused by misperceptions that tests can be reliable or valid. These misperceptions themselves may be caused, or are at least reinforced, by the use of telegraphic language that comes to be unconsciously believed as literal truth, and then unconsciously incorporated into paradigms for behavior.

Appendix A presents an addition to the on-going revision of the older AERA/APA/NCME (1985) testing standards, as proposed by the present author. One hopes that the next edition of the AERA/APA/NCME testing standards currently in development will

finally correctly treat reliability concerns.

Treatment in the Standards

The reliability standard ("A6") in the ANSI-approved program evaluation standards clearly recognize that tests are not *per se* reliable or unreliable. Rather, scores or data have varying degrees of these qualities. Standard "A6" says, "The information gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable for the intended use" (Joint Committee, 1994, p. 153).

Throughout the elaboration of this standard (pp. 153-158) references are consistently made to scores (and not tests) as being reliable. One explicitly-cited "common error" is, "C. Failing to take into account the fact that reliability of the scores provided by an instrument or procedure may fluctuate depending on how, when, and to whom the instrument or procedure is administered" (Joint Committee, 1994, p. 155).

Score/Inference Validity

Thinking about the nature of validity has steadily evolved over the last 50 years. This evolution has been traced by various authors (cf. Cronbach, 1989; Moss, 1992; Shepard, 1993).

More recent treatments of validity have emphasized that it is the inferences made from scores, and not tests, which are valid. Furthermore, recent treatments have increasingly emphasized the importance of falsification as important in evaluating validity.

The notion of time- and situation-bound validity of inferences

implies an interest in exploring the boundaries of valid score use. This interest, in turn, implies the utility of logics such as "plausible rival hypotheses" (Campbell, 1957), multitrait-multimethod evaluation of convergent and discriminant validity (Campbell & Fiske, 1959), and especially Popper's (1962) concept of falsification.

The concept of falsification requires that a theory not be deemed credible until the theory has survived serious disconfirmation efforts. As Moss (1995) explained,

> A "strong" program of construct validation requires an explicit conceptual framework, testable hypotheses deduced from it, and multiple lines of relevant evidence to test the hypotheses. Construct validation is most efficiently guided by the test of "plausible rival hypotheses" which suggests credible alternative explanations or meanings for the test score that are challenged and refuted by the evidence collected... Essentially, test validation examines the fit between the meaning of the test score and the measurement intent, whereas construct validation entails the evaluation of an entire theoretical framework. (pp. 6-7)

Treatment in the Standards

Various standards have acknowledged that inferences, and not tests, are valid or invalid. For example, the AERA/APA/NCME test standards indicate that validity evaluation requires evidence

supporting "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (AERA/APA/NCME, 1985, p. 9). Thus, it is actually the inferences from scores that are valid, and not the test, or strictly speaking even the scores across all conceivable uses.

Similarly, the personnel evaluation standards of the Joint Committee on Standards for Educational Evaluation (1988) stated that:

> Valid means that what was intended to be measured was measured. Specifically here, valid refers to the degree to which evidence supports the inferences that are drawn from the measurement instruments or procedures. Valid does not refer to the instruments or procedures themselves. Thus, a particular measure may be valid for one purpose but have little or no validity for another purpose. (p. 98)

Regarding validity, the ANSI-certified program evaluation standards (Joint Committee, 1994) state that validity "concerns the soundness or trustworthiness of the inferences that are made from the results of the information gathering process" (p. 145). Validation is "the process of compiling evidence that supports the interpretations and uses of the data and information collected using one or more of these instruments and procedures" (p. 145).

Summary

The 1994 ANSI-approved program evaluation standards dealing with reliability and validity accurately represent contemporary

views of reliability and validity within the psychometric community. As such, these standards (Joint Committee, 1994) move the field forward. One can only hope that the next in a series of editions of the AERA/APA/NCME standards will provide approximately the same treatment of reliability and validity.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA/AERA/NCME). (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), Intelligence: Measurement theory and public policy (pp. 147-171). Urbana: University of Illinois Press.

Dawis, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research:

Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Joint Committee on Standards for Educational Evaluation. (1988). The personnel evaluation standards: How to assess systems for evaluating educators. Newbury Park, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs (2nd ed.). Newbury Park, CA: SAGE.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Moss, P.A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practice, 14(2), 5-12.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Popper, K.R. (1962). Conjectures and refutations: The growth of

scientific knowledge. New York: Harper & Row.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sanders, J.W. (1994). The process of developing national standards that meet ANSI guidelines. Journal of Experimental Education, 63, 5-12.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (Vol. 19, pp. 405-450). Washington, DC: American Educational Research Association.

Thompson, B. (1991a). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1992b). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (1994a, January). It is incorrect to say "The test is

reliable": Bad language habits can contribute to incorrect or meaningless research conclusions. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Thompson, B. (1994b). The revised program evaluation standards and their correlation with evaluation use literature. Journal of Experimental Education, 63, 54-81.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

APPENDIX A

Suggested Addition of a Standard to the Draft 1996 AERA/APA/NCME Test Standards

Standard 2.23

Test users must assume some responsibility for establishing that scores are sufficiently reliable for intended uses, by thoroughly evaluating score quality from various previous reports in a detailed comparison with proposed uses, and/or by re-evaluating score quality whenever tests are administered.

Comment: Because reliability is a characteristic of scores, and not of tests, users must establish that their scores are reliable for their intended uses with their intended samples. A thoughtful comparison of score reliabilities from several previous reports, involving a detailed comparison of samples and other relevant facets of measurement, helps assure (but does not guarantee) that tests will yield reliable scores in each application. Whenever feasible, re-evaluating score quality, once data are collected, provides additional information about the scores actually being interpreted, and contributes to the body of knowledge about likely score quality in related future applications.